

Novel Core Promoter Elements and a Cognate Transcription Factor in the Divergent Unicellular Eukaryote *Trichomonas vaginalis*[▽]

Alias J. Smith,¹ Lorissa Chudnovsky,¹ Augusto Simoes-Barbosa,¹ Maria G. Delgadillo-Correa,¹
Zophonias O. Jonsson,² James A. Wohlschlegel,² and Patricia J. Johnson^{1*}

Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, California 90095-1489,¹
and Department of Biological Chemistry, University of California, Los Angeles, California 90095-1737²

Received 25 June 2010/Returned for modification 4 August 2010/Accepted 11 January 2011

A highly conserved DNA initiator (Inr) element has been the only core promoter element described in the divergent unicellular eukaryote *Trichomonas vaginalis*, although genome analyses reveal that only ~75% of protein-coding genes appear to contain an Inr. In search of another core promoter element(s), a nonredundant database containing 5' untranslated regions of expressed *T. vaginalis* genes was searched for overrepresented DNA motifs and known eukaryotic core promoter elements. In addition to identifying the Inr, two elements that lack sequence similarity to the known protein-coding gene core promoter, motif 3 (M3) and motif 5 (M5), were identified. Mutational and functional analyses demonstrate that both are novel core promoter elements. M3 [(A/G/T)(A/G)C(G/C)G(T/C)T(T/A/G)] resembles a Myb recognition element (MRE) and is bound specifically by a unique protein with a Myb-like DNA binding domain. The M5 element (CCTTT) overlaps the transcription start site and replaces the Inr as an alternative, gene-specific initiator element. Transcription specifically initiates at the second cytosine within M5, in contrast to characteristic initiation by RNA polymerase II at an adenosine. In promoters that combine M3 with either M5 or Inr, transcription initiation is regulated by the M3 motif.

The response to environmental and developmental cues is frequently orchestrated by changes in gene transcription. Processes ranging from chromatin remodeling to the precise selection of the transcription start site (TSS) must be coordinated to determine whether a gene is expressed or silent. Discrete DNA motifs play critical roles in regulating the expression of protein-coding genes, which are transcribed by RNA polymerase II in eukaryotes (8, 23, 32, 59, 65). Many motifs act as enhancer and repressor elements and are often located 100s of base pairs upstream of the TSS. Different activator and repressor complexes recognize these distal motifs and act to modulate the level of expression. Other motifs, known as core promoter elements, are in close proximity to the TSS and are required for the assembly of the RNA polymerase II preinitiation complex (PIC). Core promoter elements control the selection of the TSS and basal transcription of all protein-coding genes (23, 32, 39, 55, 59).

An array of core promoter elements that direct accurate initiation of transcription at the TSS have been identified in metazoans and include the TATA box, the initiator (Inr) element, the transcription factor_{II} B (TFIIB), recognition elements (BRE^u and BRE^d), the downstream promoter element (DPE), the motif 10 element (MTE), the downstream core element (DCE), and the X core promoter elements (XCPEs; XCPE1 and XCPE2) (1, 5, 16, 21, 32, 33, 37, 50, 58, 66). These can function independently, as in the case of the TATA box,

Inr, and XCPEs, or in varied combinations to direct basal transcription.

By and large, what we know about eukaryotic core promoters and the transcription factors that specifically recognize them and that act in concert to direct initiation of transcription comes from studies conducted in fungi, plant, and animals (25, 59). In contrast, the properties governing gene expression in the broad range of divergent unicellular eukaryotes known as protists (3), including *Trichomonas vaginalis*, remain poorly understood. *T. vaginalis* is a parasitic protist responsible for the most common nonviral sexually transmitted infection worldwide: trichomoniasis (52, 57, 69, 72). While infections may be asymptomatic, *T. vaginalis* causes urethritis in men and often causes vaginitis in women (20, 52), leading to an increased risk of prostate (61, 63) and cervical (74, 75) cancers, respectively. Trichomoniasis is also a risk factor for the acquisition of HIV (43, 47, 60, 67). *T. vaginalis* belongs to the phylum Parabasalia within the kingdom Excavata, once argued to be among the earliest-diverging eukaryotic lineages (9, 24, 35). More recent phylogenetic analyses have challenged the early divergence of the Parabasalia but still clearly distinguish this group as being highly divergent from other eukaryotic lineages (3, 34).

Our previous transcription studies on *T. vaginalis* gene transcription have revealed several novel key properties, as well as conserved features shared between this divergent microbe and metazoans (36, 40, 42, 54, 56). Like that found in metazoans, *T. vaginalis* protein-coding genes are composed of bipartite promoters with distal elements that regulate the level of transcription and a core promoter that directs accurate transcription initiation (28, 40, 41, 50). To date, the Inr element is the only core promoter element that has been described in *T. vaginalis*. This Inr element, which contains and surrounds the TSS, is structurally and functionally equivalent to the metazoan Inr

* Corresponding author. Mailing address: Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, CA 90095-1489. Phone: (310) 825-4870. Fax: (310) 206-5231. E-mail: johnsonp@ucla.edu.

[▽] Published ahead of print on 18 January 2011.

element (41, 42, 54). However, unlike the metazoan Inr, which is recognized by the TATA box-binding protein (TBP)-associated factor 1 (TAF1)/TAF2 complex, a component of the general transcription factor TFIID complex (10, 59), the *T. vaginalis* Inr is bound by a novel 39-kDa transcription factor (initiator binding protein 39 [IBP39]) not present in any other organism. IBP39 binds the *T. vaginalis* Inr and mediates transcription initiation (42, 56). Studies indicate that IBP39 also interacts with the C-terminal domain of the large subunit of RNA polymerase II and is proposed to aid in the recruitment of RNA polymerase II to the TSS (36, 56).

In addition to the Inr element, metazoans also utilize the TATA box to direct accurate transcription initiation (23, 59). However, no *T. vaginalis* genes have been shown to contain functional TATA-like motifs in the location in which they are found in metazoan genes (~25 to 30 bp upstream of the TSS). In this study we have examined a database of sequences upstream of expressed *T. vaginalis* genes to determine whether this microbe uses TATA-like elements to drive transcription and to identify other core promoter motifs that might be present. Candidates found using this bioinformatics approach were then subjected to analysis by use of the following criteria to define core promoters: (i) the DNA motif should be overrepresented, relative to random chance, in the 5' untranslated regions (UTRs) of many protein-coding genes, (ii) its location should be fixed relative to the transcription start site, and (iii) mutations within the motif should affect transcriptional activity (31). Two novel eukaryotic core promoter elements, motif 3 (M3) and motif 5 (M5), were found using this approach, and a third was identified during the analysis of core promoter architecture. M3 resembles a Myb recognition element (MRE) found in the distal promoter region of genes in other eukaryotes but has not previously been shown to serve as a core promoter of protein-coding genes. A specific M3 binding protein (M3BP) was also identified and shown to be a novel protein with a Myb-like DNA binding domain. The M5 element overlaps the TSS and is a gene family-specific alternative initiator element. Transcription initiating within M5 is unusual, as the TSS is marked by a cytosine. Transcription from both the M5 and the Inr were found to be regulated by M3. These studies greatly extend our knowledge of the characteristics of core promoters in this divergent eukaryote.

MATERIALS AND METHODS

Creation of upstream region database. A nonredundant upstream region database (URDB) containing the 5' UTRs of protein-coding genes with evidence for expression was created by first identifying and downloading predicted protein-coding gene sequences of the *T. vaginalis* genome (www.Trichdb.org). Expressed sequence tag (EST) data consisting of 20,002 sequences were then downloaded from GenBank, and BLASTn analysis was used to identify the protein-coding genes with expression evidence. Any gene that did not match 100% to an EST was eliminated. Multiple genes that were or nearly were 100% identical were then compared over 30 bp upstream and downstream of the translation start site using CLEANUP software (22). Only one member within a group of genes that were at least 95% identical was retained to create the nonredundant URDB used in all further analyses.

DNA motif searches. A Perl script was written to search the URDB sequences 60 bp upstream to 40 bp downstream of the predicted start of translation for DNA motifs similar to known eukaryotic core promoters. The output of the Perl script included the frequency of each motif and its position within individual URDB sequences. The DNA motifs included in the search were the metazoan BRE^a [(G/C)(G/C)(G/A)CGCC], archaea BRE^b [(C/A)N(A/T)AA(A/T)], metazoan BRE^d [(G/A)T(T/G/A)(T/G)(T/G)(T/G)], metazoan TATA box [TATA

(A/T)A], archaea TATA box [TTTA(A/T)ATA], Inr [(T/C/T)(C/T)A(C/T)(T/A)], DPE [(G/A)G(A/T)(A/T)(G/A/C)], MTE [(G/C)A(G/A)C(G/C)(G/C/A)AACG(G/C)], XCEP1 [(G/A/T)(G/C)G(T/C)GG(G/A)A(G/C)(A/C)], and XCEP2 [(A/C/G)C(C/T)C(G/A)TT(G/A)C(C/A)(C/T)] (1, 5, 32, 37, 50, 66).

A local installment of the MEME program (version 3.5.7) was used to search for overrepresented DNA motifs using a motif window of 6 to 8 or 6 to 10 bp and zero or on occurrence per sequence (2). The results obtained using each search window were similar. MEME uses an iterative expectation-maximization algorithm to identify conserved motifs, yielding ungapped blocks of sequences that are represented by weight matrix models. Motifs that contain gaps were identified as individual motifs.

Mapping of TSSs (RNA ligase-mediated [RLM] 5' rapid amplification of cDNA ends [RACE]). Total RNA from *T. vaginalis* strain G3 was isolated using Trizol LS (Invitrogen). The RNA was LiCl precipitated, ethanol precipitated, and DNase treated (Ambion). mRNA was then isolated using a PolyAtract mRNA isolation system III (Promega), following the manufacturer's recommendations, and treated sequentially with calf intestine alkaline phosphatase (CIP) and tobacco acid pyrophosphatase (TAP) (Invitrogen). An RNA adapter was ligated onto the 5' end of the CIP/TAP-treated mRNA and then reverse transcribed using Superscript III (Invitrogen) and oligo(dT) (Invitrogen). The resulting pool of cDNA was used in PCRs with a manufacturer-provided RNA adapter-specific forward primer and gene-specific reverse primers (data not shown). The PCR products were cloned into the Strataclean PCR cloning vector and sequenced to determine the 5' nucleotide of the target cDNAs (Genewiz).

Plasmid construction. The alpha-succinyl coenzyme A synthetase-chloramphenicol acetyltransferase (alpha-SCS-CAT) construct described by Delgadillo et al. (12) was used to construct plasmids that contained M3/Inr, M3/M5, or M5 containing upstream regions preceding the *CAT* reporter genes. Forward primers encoding a SacI restriction site and reverse primers containing an NdeI restriction site were designed to amplify ~500 bp of target upstream regions. The primers used for plasmid construction are listed below, and the restriction sites are underlined: 5'-GAGCTCGAACCATTTCAGGATTATATTAAATTAT T-3' for TVAG_359800_F, 5'-CATATGTGGTACAAAGGACAGCCATAAC CG-3' for TVAG_359800_R, 5'-GAGCTCCTTCTCTCGAATTGTAAAGAAA ATCGG-3' for TVAG_055940_F, 5'-CATATGTTTTTCAAAAGGCCAACCAA ACC-3' for TVAG_055940_R, 5'-GAGCTCCCAAGACACCTATCTCGTCGGG AATC-3' for TVAG_380910_F, 5'-CATATGTAAAAAAGGCCAACCAAAC CCTC-3' for TVAG_380910_R, 5'-GAGCTCCTCCGGATTGTAATTCTGTT GCTAC-3' for TVAG_333620_F, 5'-CATATGCCAAAAGTGAATTACAA CCGTAC-3' for TVAG_333620_R, 5'-GAGCTCCACCATCCACATCATC TAATTGGTCTCTC-3' for TVAG_090090_F, and 5'-CATATGTTTCGATAAAA GTAAAAACCGTCATATTTTCATCG-3' for TVAG_090090_R. The amplified upstream regions were subcloned into the alpha-SCS-CAT expression plasmid, replacing the alpha-SCS upstream region.

Targeted mutations were introduced using the QuikChange site-directed mutagenesis method (Stratagene). Sequence-specific primers encoding the desired mutation flanked by 15 to 20 nucleotides (nt) of complementary sequence were used with wild-type M3/Inr-CAT, M3/M5-CAT, or M5-CAT plasmids as templates. All plasmids were sequenced to confirm the presence of the desired mutation (Genewiz).

Quantitative real-time PCR (qRT-PCR). Total RNA was isolated from *T. vaginalis* strain T1 populations transfected with wild-type or mutant 5' UTR-CAT plasmids using Trizol LS (Invitrogen). The RNA was LiCl precipitated, ethanol precipitated, DNase treated (Ambion), and reverse transcribed using Superscript III and oligo(dT) (Invitrogen). Real-time PCRs were performed with the resulting cDNA, brilliant SYBR green QPCR master mix (Stratagene), and primers for the *CAT* reporter gene (5'-GAAAGCGGTGAGCTGGTGATAT G-3' and 5'-ACTGGTGAAACTCACCCAGGGATT-3') or the neomycin gene (5'-ACAGACAATCGGCTGCTCTGATG-3') and 5'-ACCATGATATTCGGC AAGCAGGCA-3'). cDNA reactions without reverse transcriptase were used as negative controls. All wild-type and mutant *CAT* reactions were normalized to neomycin expression.

Preparation of *T. vaginalis* crude nuclear extract. *T. vaginalis* strain G3 was harvested by centrifugation and washed in 1× phosphate-buffered saline (PBS), followed by one wash in buffer A (10 mM HEPES, pH 7.9, 1.5 mM MgCl₂, 10 mM KCl, 0.2 mM phenylmethylsulfonyl fluoride [PMSF], 0.5 mM dithiothreitol [DTT], 10 µg/ml leupeptin, 50 µg/ml *N*-α-p-tosyl-L-lysine chloromethyl ketone [TLCK]). The cells were then resuspended in buffer A and lysed. The lysate was cleared by centrifugation and the pellet was resuspended in 20 mM HEPES, pH 7.9, 1.5 mM MgCl₂, 400 mM KCl, 25% glycerol, 0.2 mM PMSF, 0.5 mM DTT, 10 µg/ml leupeptin, and 50 µg/ml TLCK. The supernatant was dialyzed against a 50× volume of buffer E (20 mM HEPES, pH 7.9, 0.2 mM EDTA, 100 mM KCl, 20% glycerol, 0.2 mM PMSF, 0.5 mM DTT, 10 µg/ml leupeptin, 50 µg/ml TLCK).

overnight. All steps were performed at 4°C. The resulting crude nuclear extract was either used fresh or immediately stored at -80°C.

Electrophoretic mobility shift assay (EMSA). Nucleotide probes containing M3 consensus sequences (see Fig. 6C) were prepared by annealing complementary oligonucleotides and end labeling the resulting double-stranded DNA with gamma-³²P and T4 polynucleotide kinase (New England BioLabs). The labeled probes were gel purified on a 6% polyacrylamide gel. Unlabeled annealed primers encoding wild type or mutant M3 sequences were also prepared for use in competition assays. Each binding reaction mixture contained either 20 µg of *T. vaginalis* crude nuclear extract or 20 ng of recombinant M3BP (rM3BP), 10,000 cpm of labeled probe, 500 ng poly(dI-dC), 20 mM HEPES-KOH (pH 7.9), 100 mM KCl, 1 mM DTT, 1 mM EDTA, 0.01% NP-40, and 15% glycerol. For the competition assays, 100× molar excess of the relevant unlabeled probe was incubated in the binding reaction mixture prior to the addition of labeled probe.

Identification and isolation of M3BP. M3BP was isolated using DNA affinity chromatography as described previously (46). The wild-type (M3-1 wt) and mutant (M3-1 mut3) M3 sequences used are shown in Fig. 6C. The double-strand oligonucleotides were biotinylated and fixed to streptavidin-Sepharose beads. Gravity columns were created using the probe-streptavidin-Sepharose. One milligram of crude *T. vaginalis* nuclear extract (NE), diluted in the EMSA binding buffer described above, was precleared with probe M3-1 mut1 by incubating the mixture at room temperature for 20 min with rotation. The precleared NE was then split into two equal fractions. One was applied to a wild-type M3-1 probe gravity column, and the second was applied to an M3-1 mut1 gravity column. The columns were washed with EMSA binding buffer, followed by subsequent washes with EMSA binding buffer containing increasing concentrations of KCl, ranging from 100 mM to 500 mM KCl. The crude NE, precleared NE, flowthrough, and wash fractions from each column were divided in half. One half was concentrated and used to track the M3 DNA binding activity via EMSA using gamma-³²P-labeled wild-type probe M3-1. The other half of the 400 mM KCl fractions from both columns was dried and submitted for multidimensional protein identification technology (MudPIT) mass spectrometry to identify the putative M3BP.

MudPIT mass spectrometry analysis. Affinity-purified samples were precipitated by the addition of trichloroacetic acid to a final concentration of 10%, washed with ice-cold acetone, and then resuspended in digestion buffer (100 mM Tris-HCl, pH 8.5, 8 M urea). Proteolytic digestion was performed by the sequential addition of Lys-C and trypsin proteases and analyzed using shotgun proteomic methods on an LTQ-OrbitrapXL mass spectrometer (ThermoFisher) as previously described (18, 68, 70). Computational analysis of mass spectra was performed using the SEQUEST and DTASelect algorithms (15, 64). Data were filtered so that a protein identification required two unique peptide identifications per protein using a peptide-level false-positive rate of 5%, as determined using a decoy database strategy (14).

Cloning, expression, and purification of M3BP. The gene encoding the Myb-like protein was cloned into the *Escherichia coli* expression vector pET-200D (Invitrogen), using the primers 5'-CACCATATGCAAACCTCCATGTCTAAC G-3' and 5'-TTAATGGTGTGATGGTGTTCCTGAGGTAAGATCAA TGTATTGAGC-3'. The reverse primer encoded a six-histidine tag. Induced bacterial lysates were sequentially purified over a Mono S cation-exchange column (Bio-Rad), followed by purification over an NiSO₄-charged chelating Sepharose column (Amersham), according to the manufacturer's recommendations, to remove additional contaminants and further enrich for full-length rM3BP. The resulting C-terminal His-tagged rM3BP was assayed via EMSA, as described above.

Chromatin immunoprecipitation analyses. A *T. vaginalis* transfectant expressing the 40S ribosomal protein S15-2 was grown overnight to a density of ~4 × 10⁶ cells/ml, and formaldehyde was added to a final concentration of 1%, followed by incubation on a rotator for 20 min at room temperature. Cross-linking was stopped by the addition of 125 mM glycine, followed by a 5-min incubation at room temperature. Cells were harvested and washed in 20 ml cold PBS plus 125 mM glycine and protease inhibitors, split into 500-µl aliquots, and sonicated to shear the DNA to an average of ~400 nucleotides. The extract was spun at 13,000 rpm for 15 min at 4°C, and the supernatant was subjected to immunoprecipitation (IP). Two IPs, each containing 800 µl of the extract, were performed. One contained 20 µl of IgG purified from anti-M3BP antisera using protein A-Sepharose, and the other had no antibody (negative control). The IP reaction was done overnight at 4°C on a rotator. Thirty-five microliters of protein A beads was then added to each reaction mixture, which was then incubated for 2 h at 4°C. The beads were washed in a low-salt wash (1% Triton X-100, 1 mM EDTA, 50 mM HEPES, pH 7.5, 150 mM NaCl, 0.1% deoxycholate), followed by a high-salt wash (1% Triton X-100, 1 mM EDTA, 50 mM HEPES, pH 7.5, 500 mM NaCl, 0.1% deoxycholate), an LiCl wash (250 mM LiCl, 0.5% NP-40, 0.5%

deoxycholate, 1 mM EDTA, 10 mM Tris, pH 8), and a TE wash (10 mM Tris, pH 8, 1 mM EDTA). All washes were done in 1 ml for 5 min at room temperature on a rotator. The beads were suspended in 100 µl elution buffer (1% SDS, 50 mM Tris, pH 8, 10 mM EDTA) and incubated for 10 min at 65°C. After elution, the beads were spun and the supernatants were incubated overnight at 65°C to reverse cross-linking. The samples were then treated with proteinase K, and DNA was recovered by phenol-chloroform extraction and ethanol precipitation. PCR was done using the recovered DNA with an annealing temperature of 48°C, *Taq* polymerase, and 2.4 mM MgCl₂. The primers used were GCCTGGTAGT ATATGAATCACC (forward) and GGATAACAACCTCTTGTGTCC (reverse) for the 40S ribosomal protein S15-2 and GTTCGCTAACTACGATCTTC (forward) and AGTAAGAAGAATTTGCAATCCG (reverse) for the noncoding DNA on contig DS113477 (www.trichDB). The PCR products were analyzed on a 2% agarose gel.

RESULTS

Identification of overrepresented motifs in UTRs of *T. vaginalis* protein-coding genes. Only one core promoter has been shown to direct transcription initiation of protein-coding genes in the unicellular eukaryote *T. vaginalis*: the Inr [(T/C/A)(C)(A₍₊₁₎)(T/C/A)(T/A), where (+1) denotes the TSS] (41, 42, 56). To search for additional potential core promoter elements, we created a *T. vaginalis* protein-coding gene URDB selected from 20,002 ESTs (www.trichdb.org). Given the high frequency of multigene families in *T. vaginalis*, many of which appear to have undergone recent expansion (7), the CLEANUP program (22) was used to identify groups of sequences ≥95% identical within 30 bp upstream and downstream of the translation start site. Only one member of each group was retained in the URDB to avoid overrepresentation of 5' UTRs reflecting recent expansion and not conservation. The resulting nonredundant data set contained 1,647 5' UTRs from expressed genes. These UTRs have a 23.9% G+C content, comparable to the 28.8% G+C content reported for *T. vaginalis* intergenic regions (7).

Approximately 82% of the UTRs within the URDB were found to contain the Inr consensus sequence in the previously determined location for this promoter element: 6 to 20 bp upstream of the translation start site ATG (41). As a first step toward identifying potential DNA motifs that may act as core promoters in the 18% of the genes that do not contain a putative Inr, we used the MEME program (2) to search the entire URDB for overrepresented motifs. On the basis of the close proximity of functional Inrs to the ATG translation start site in *T. vaginalis* genes (41), the 60 bp upstream of the ATG was examined for motifs with a maximum width of 8 bp. The top five statistically significant motifs identified are shown in Table 1. The distribution of the five motifs relative to the ATG start codon was assessed, and all except motif 2 were found to be positionally conserved (Fig. 1). On the basis of the random distribution of motif 2 and the strong association of motif 4 with the translation start site, these motifs were excluded from further analysis.

The most statistically significant motif identified by MEME, motif 1 (M1), contains a *T. vaginalis* Inr core element [TCA (C/T)T] at its 5' end (Table 1). As predicted, the majority (~65%) of identified M1s are located 5 to 20 bp upstream of the ATG (41). These data underscore the importance of the Inr in directing transcription initiation and validate this method as a way to uncover overrepresented DNA motifs. As we have previously rigorously analyzed the core motif of M1

TABLE 1. The five most significant motifs identified by the MEME algorithm

Motif	Pictogram	IUPAC consensus sequence	No.	%	e value
1		TCAYTTTT	562	34.1	6.3e-291
2		AAAGTGAC	189	11.5	2.5e-134
3		RRCSGTTD	197	11.9	5.7e-117
4		MAAAAWTK	452	27.4	1.1 e-54
5		GDCCTTTY	73	4.4	4.3e-49

(40–42, 54), the motif has not been subjected to further biochemical analyses here.

M3 is present in 11.9% of sequences and has a peak location distribution at positions –20 to –25 relative to the translation start site. M3 is present in the 5' UTRs of many unrelated gene families. Three of the 8 nt of M3 exhibit 100% conservation, and the first five positions of the consensus sequence [(A/G/T)(A/G)C(G/C)G(T/C)T(T/A/G)] are similar to the last five positions of the mammalian MRE [Py-AAC(T/G)G, where Py is pyrimidine].

The consensus sequence of the remaining motif identified by MEME, M5 [G (G/A/T)CCTT(T/C)(T/C)], does not resemble any known eukaryotic RNA polymerase II promoter element. This motif has the most striking positional conservation, with ~80% being located at positions –10 to –15 relative to the translation start site (Fig. 1), the position where functional Inr

elements are typically located (41). Interestingly, 85.7% of M5-containing sequences precede ribosomal protein genes, a gene family predicted to have 422 members in *T. vaginalis* (www.trichdb.org). To determine the percentage of ribosomal protein genes contain an M5-like element, a data set of the 5' UTRs of the 422 sequences was searched using MEME as described above, and 86.4% of the sequences were found to contain an M5 in its peak location. Notably, 46.7% of these M5-containing UTRs also have a positionally conserved M3, while only ~6% contain a potentially functional Inr element.

As no previously described eukaryotic core promoter elements were identified by the MEME program, with the exception of the Inr element, we used custom Perl scripts to directly search the URDB for the core promoter elements: TATA box, TFIIB, BRE^a, BRE^d, DPE, MTE, DCE, XCPE1, and XCPE2 (1, 5, 32, 37, 50, 66). Although all are represented, with the exception of MTE and XCPE2, little to no positional conservation is observed. Hence, the only known eukaryotic core promoter motif that appears to be present in *T. vaginalis* protein-coding genes is the previously described Inr (40, 41, 55).

M3 and M5 are found in conserved positions relative to TSS. As core promoter elements are typically conserved in position relative to the TSS, we used RLM 5' RACE to determine the relationship, if any, between the TSS and M3 and M5 (Fig. 2). All TSSs were found to be located 7 to 15 bp upstream of the translation start site, within the range of locations of previously reported Inr-driven initiation of transcription (Fig. 2). In UTRs that contain both Inr and M3, the TSS was mapped primarily at the adenosine within the Inr at a distance 11 to 17 bp downstream of the M3 (Fig. 2A). A survey of the URBD reveals that this relative location of M3 and Inr is conserved in ~33% of UTRs that contain both elements.

We next examined 5' UTRs that do not contain the Inr. We first mapped the TSS for 14 UTRs that contain both M3 and M5 (Fig. 2B) and 6 that contain only the M5 element (Fig. 2C)

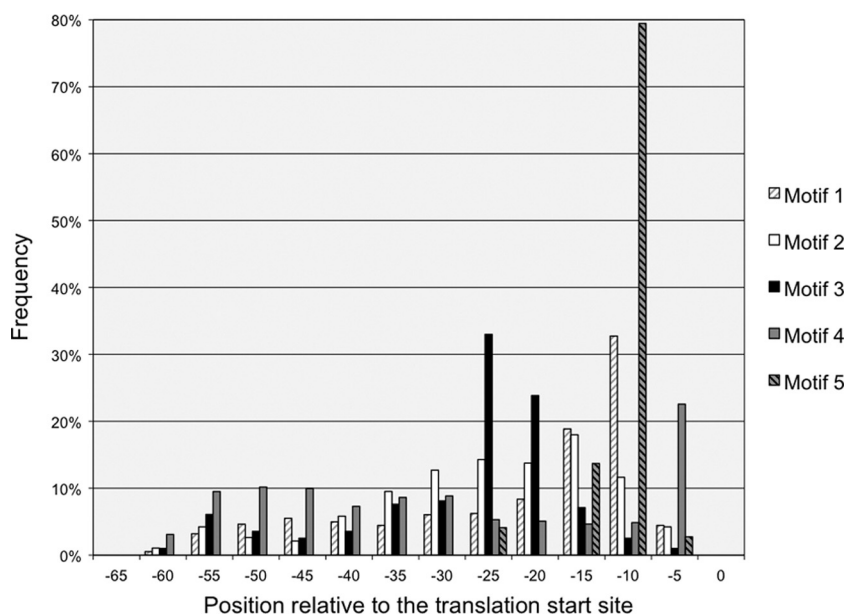


FIG. 1. Location distribution of the top five significant MEME-identified DNA motifs relative to the translation start site. The translation start site is represented by 0, and the position of the first nucleotide of each motif, as shown in Table 1, was binned in 5-bp intervals.

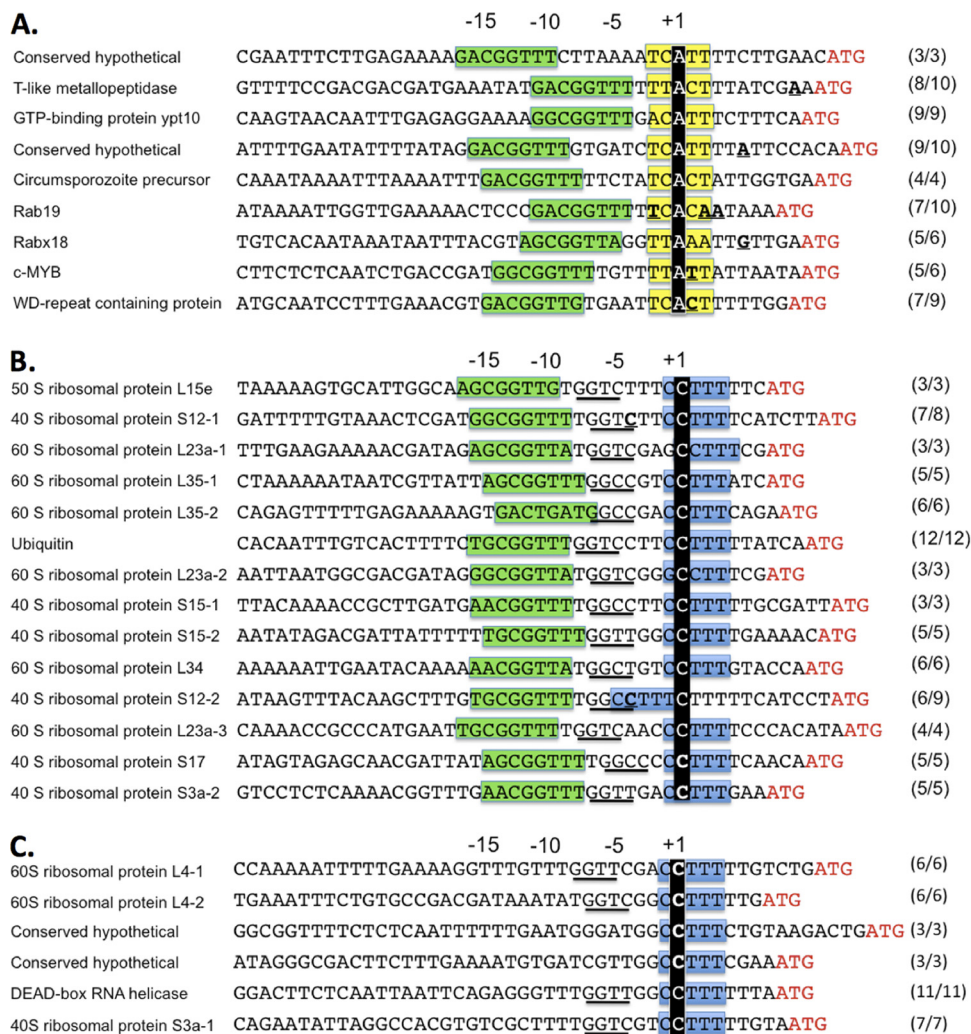


FIG. 2. Positions of motif 3 (green), motif 5 (blue), and the Inr element (yellow) relative to the transcription start site, determined by RLM 5' RACE. The ATG translation initiation codon is shown in red. The transcription start site is denoted +1 and is highlighted by the black bar. The frequency at which each highlighted TSS was mapped is listed on the right. Additional TSSs mapped for a given sequence are represented in boldface. (A) UTRs containing M3 and Inr; (B) UTRs containing M3 and M5; (C) UTRs containing only M5. The non-MEME-identified GG(T/C)(T/C) motif is underlined in panels B and C.

and found that with two exceptions the TSS mapped within M5, at the second C of the conserved dinucleotide $C_{(-1)}-C_{(+1)}$. The location of M3 in M3-M5 UTRs is conserved and located -14 to -17 nt upstream of the TSS. A closer inspection of the mapped TSSs revealed the strict conservation of the $CC_{(+1)}TTT$ motif at the core of the 8-nt motif defined by MEME with variation at other positions, which led us to redefine M5 as CCTTT (highlighted in Fig. 2B and C). This redefined M5 is present in ~12% of the 60-nt UTRs in the URDB, with ~71% being located between -5 and -15 nt relative to the translation start site. The mapping of the TSS to the M5 in UTRs that lack an Inr and the overlapping location of these 2 elements implicate M5 as an alternate *T. vaginalis* initiator element. Interestingly, M5 is devoid of adenosines and instead uses a noncanonical cytosine to mark the TSS. To confirm the TSS mapping data obtained using RLM 5' RACE, conventional primer extension analyses were also conducted on six genes, and the same TSS was found in all cases (data not shown).

Further analysis of the TSS mapping data revealed an additional 4-nt DNA motif unidentified by our MEME analyses which is underlined in Fig. 2B and C. This GG(T/C)(T/C) motif precedes M5 in all shown M3-M5 5' UTRs and in 4 out of 6 M5-only UTRs and is located 6 to 8 bp upstream of the TSS. Reanalysis of the URDB shows that ~75% of M3-M5 and ~17% of M5-alone 5' UTRs have this motif conserved within 15 bp upstream of M5.

The M3 and GG(T/C)(T/C) motifs affect transcriptional activity. There is no reliable *in vitro* transcription system for examining the expression of *T. vaginalis* genes; therefore, we have used *in vivo* assays to determine whether M3, M5, and the GG(T/C)(T/C) motif affect transcriptional activity. Approximately 500 bp of selected target 5' UTRs was cloned upstream of a CAT reporter gene in a *T. vaginalis* expression vector (41), and mutations within the motifs were introduced by site-directed mutagenesis. Wild-type and mutant constructs were then introduced into *T. vaginalis*, and the transcriptional activ-

ity of the *CAT* reporter gene was assayed by qRT-PCR. All constructs contained a neomycin phosphotransferase gene that was also subjected to qRT-PCR to correct for possible variation in the number of plasmid constructs in transfectants being compared. The 5' UTRs of 2 genes containing both the M3 and M5 motifs were selected for analysis (60S ribosomal protein L34 [TVAG_359800] and 40S ribosomal protein S15-2 [TVAG_055940]) (Fig. 3A and B). The TSS for each of these maps to the second cytosine of M5, and each contains the GG(T/C)(T/C) motif between the M3 and M5 motifs. The 5' UTRs of a WD repeat-containing protein (TVAG_333620) and T-like metalloproteinase (TVAG_090090) were chosen as representatives of M3-Inr-containing 5' UTRs where the TSS maps to the canonical adenosine within the Inr (Fig. 4A to D).

A triple-base-pair mutation (GTT to TAA) or three single mutations at the strictly conserved C, G, or T in the M3 element (Table 1) upstream of the 60S ribosomal protein were found to nearly abolish transcriptional activity, reducing it to ~0.5, ~0.6, ~0.1, and ~0.1% relative to that of the wild type, respectively (Fig. 3A, mut1, mut2, mut3, and mut4). The same single mutations in M3 upstream of the 40S ribosomal protein likewise dramatically decreased transcriptional activity (Fig. 3B, mut12, mut13, and mut14). These single-base-pair mutations also decreased transcriptional activity in the context of the M3-Inr promoters, although not as dramatically as that observed for M3-M5 promoters (Fig. 4A, mut27 and mut28, and Fig. 4C, mut30, mut31, and mut32), with activity ranging from ~5.5 to ~34% of that of wild-type promoters. Together, these data demonstrate that, when it is present, M3 modulates the level of transcription driven from either the Inr or M5 motif.

We next asked if the GG(T/C)(T/C) motif located between M3 and M5 affects transcriptional activity by mutating each position of the motif in the two M3/M5-containing UTRs (Fig. 3A and B). The first two positions were changed to A residues and the second two positions were mutated to G residues. All four mutations in each M3-M5 promoter dramatically decreased transcriptional activity to between ~0.02 and ~4% (Fig. 3A, mut8, mut9, mut10, and mut11) or ~10 and ~14% (Fig. 3B, mut18, mut19, mut20, and mut21) of that of the wild type. These analyses demonstrate that, similar to what was observed for M3, the GG(T/C)(T/C) motif is a core promoter element.

M5 functions as an initiator element. M5 surrounds and contains the TSS, indicating that it is likely to function as an alternative initiator element. To test whether mutations within M5 alter transcriptional activity, we introduced single-base-pair mutations in the M5 of the 60S ribosomal protein gene, changing the C to G at position +1, C to G at position -1, and T to G at position +3 relative to the TSS (Fig. 3A). Mutating the C that marks the TSS (mut5) had the largest effect, decreasing transcriptional activity to ~15% of that of the wild type. Mutating the C at the -1 position (mut6) decreased the activity to ~47% of that of the wild type. There was no significant difference in transcriptional activity when the T at position +3 (mut7) was mutated. To confirm and extend upon these results, M5 mutations were analyzed in the context of another M3/M5 promoter (Fig. 3B). The C at position +1 was mutated to an A, a G, or a T, and transcriptional activity was

found to be reduced to ~15 to ~35% of that of the wild type promoter (mut15, mut16, and mut17, respectively).

We next chose the DEAD-box RNA helicase (TVAG_380910) to examine a 5' UTR that lacks an M3 but that does contain the M5 and GG(T/C)(T/C) motifs with the TSS mapping to the second C within M5 (Fig. 3C). We made single-base mutations at each position of the M5 motif (CCTTT) within this 5' UTR, mutating each to a G (Fig. 3C). Less than 15% activity was observed when either the C at position -1 or the C at position +1 was mutated (mut22 and mut23), whereas considerably higher activity (~35%) was found when the T at position +2 was mutated (mut24). As also seen when the T at position +3 was mutated in the UTR of the 60S ribosomal protein gene (Fig. 3A, mut7), mutation of the element at this position in the DEAD-box RNA helicase promoter did not significantly affect activity (Fig. 3C mut25). Similarly, only a modest reduction in activity was observed when the T at position +4 was mutated (mut26). These data demonstrate that mutation of the TSS and the immediately surrounding nucleotides significantly reduces transcriptional activity, thus identifying M5 as an alternative initiator element. Notably, although they are strictly conserved, the last two T residues in the M5 motif do not directly contribute to the strength of transcriptional activity.

Mutations within M5 alter the TSS selection. To determine whether mutations in M3 and M5 alter the TSS selection, RLM 5' RACE was used to map the 5' ends of RNAs derived from selected mutants. With a single exception, we found that mutations at positions -1, +1, and +2 within M5 alter the selection of the TSS, whereas mutations in M3 do not alter the TSSs mapped within either M5 or Inr (Fig. 3A and B; Fig. 4A and C; curved arrows denote the TSS). When the M5 element is mutated at either the C at position -1 or the C at position +1 to G or T, the TSS is shifted downstream to A residues (Fig. 3A, B, and C, mut5, mut6, mut16, mut17, mut22, and mut23). In most cases, these adenosines are found within Inr-like elements similar to ones we have previously shown to function as Inr elements with reduced activity (41). The only tested mutation at the C at position -1 or +1 in M5 that does not lead to a change in the TSS occurs when the C at position +1 is mutated to an A (Fig. 3B, mut15). This changes the M5 into a degenerate, low-activity Inr (41), which likely accounts for the 80% reduction in activity and no shift in the TSS. Although 6/10 TSSs mapped to the C at position +1 when the T at position +2 was mutated, 4/10 TSSs were shifted to A residues downstream of M5 (Fig. 3C, mut24). In contrast, no shift in the TSS was observed when the T residues at positions +3 and +4 were mutated (mut25 and mut26). Thus, the two T residues at the 3' end of M5, which were shown above to have no effect on the strength of transcriptional activity, also do not affect the selection of the TSS. Given the conservation of these TT residues and the role of poly(T) in destabilizing double-stranded DNA (dsDNA), these nucleotides may be important for facilitating melting of the dsDNA to form an open complex for the initiation of transcription while not significantly affecting the transcriptional activity of the promoter or potential DNA-protein interactions (73).

The distance between motifs affects transcriptional activity and TSS selection. To directly test whether the observed conservation of the distance between M3-M5 and M3-Inr promot-

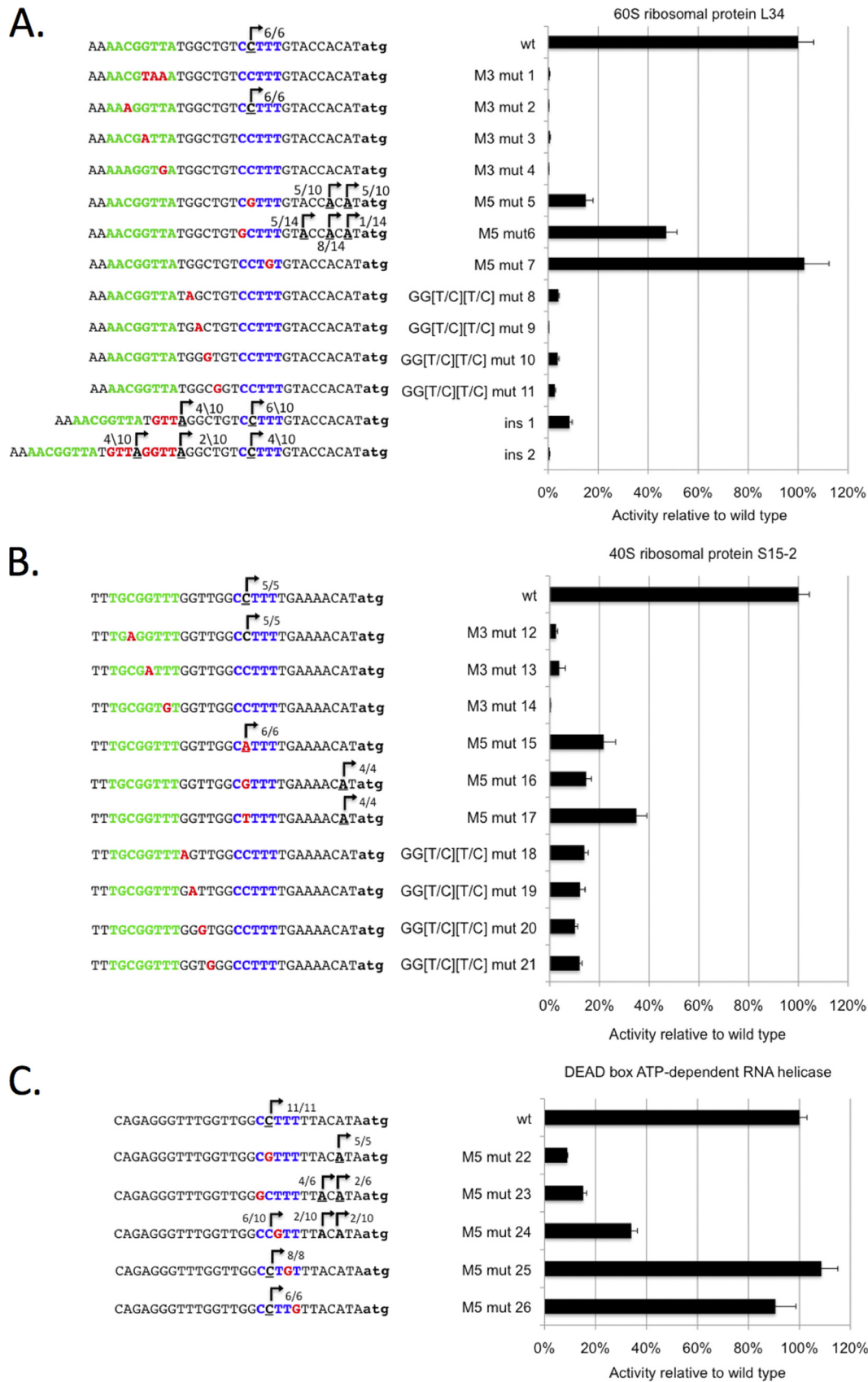


FIG. 3. Transcriptional activity of genes with M3-M5 and M5 promoters. The sequence ~500 bp upstream of the 60S ribosomal protein L34 (A), 40S ribosomal protein S15-2 (B), or DEAD-box ATP-dependent RNA helicase (C) was cloned 5' of a *CAT* reporter gene, and

ers (Fig. 1, 2A and B) plays a role in directing transcription, we introduced insertions and deletions in reporter gene constructs containing these promoters. When either 4 or 9 bp was inserted between M3 and the GG(T/C)(T/C) motif in an M3-M5 promoter, transcriptional activity was reduced to ~8.5 and ~0.5% of that of the wild type, respectively (Fig. 3A, ins 1 and ins 2), demonstrating that the distance between M3 and the downstream GG(T/C)(T/C) and M5 elements is critical for transcriptional activity. The distance between the M3 and Inr in promoters containing both motifs was also found to affect transcriptional activity (Fig. 4A, ins 3 and ins 4), however not as dramatically as in M3-M5 promoters (~50%). This is likely attributed to the ability of the Inr to function independently as a promoter (41).

Increasing the distance between M3 and M5 also altered the TSS, moving it to A residues upstream of the TSS with a C at position +1 for 40 to 60% of mapped mRNAs (Fig. 3A, ins1 and ins2); the TSSs of the remaining mRNAs mapped to the wild-type C at position +1 in the M5 motif. It is notable that the altered TSSs are located at +12 and +17 nt downstream of M3, consistent with the 11 to 17 nt found between M3 and TSS in wild-type promoters. These data strongly indicate that M3 is involved in selecting the TSS in M5-containing promoters and that the optimal distance between M3 and the TSS is 12 to 17 nt. Transcription initiating in the M5 motif is not strictly dependent on the conservation of M3 at this location, however. This implies that an increased distance between these elements can support M3-directed transcription or that M5 is capable of directing transcription in the absence of M3, albeit at a highly reduced level.

Increasing the distance between M3 and Inr by 5 nt in promoters containing both elements also creates an alternative TSS at an A 15 bp downstream of M3 (Fig. 4A, ins 3). This TSS was found in ~43% of mapped mRNAs and is within the 11 to 17 bp observed for M3-TSS spacing. However, when the distance between the 2 promoters is increased by an additional 5 nt, no alternative TSS is created and transcription initiates exclusively at the Inr (Fig. 4A, ins 4).

The effect of mutations in M3 has a more dramatic effect on transcription driven from the Inr when only 1 bp separates the two elements (Fig. 4C, T-like metallopeptidase promoter) than when 5 bp separates them (Fig. 4A, WD-repeat-containing protein promoter). To further test the effect of M3/Inr spacing, we deleted 4 of the 5 bp separating M3 and Inr in the WD-repeat-containing protein promoter (Fig. 4B) and conversely inserted 4 bp between these elements in the T-like metallopeptidase promoter (Fig. 4D). Deletion of the 4 bp increased transcription ~5-fold (Fig. 4B, del). This increase in transcription is dependent on M3, as introducing a mutation in M3 in this construct decreased transcription to ~15% of that of the wild type (Fig. 4B, mut29 del). Similarly, increasing the spacing between M3 and Inr by 4 bp in the T-like metallopeptidase

promoter decreased activity to ~50% of the wild-type levels (Fig. 4D, ins 5), and mutation of M3 in this context further decreased activity, demonstrating direct modulation of expression by M3 (mut33 ins 5). These data demonstrate that M3 and Inr work synergistically, with the distance between these elements substantially affecting the level of transcription, without altering the TSS (Fig. 4B and D).

A *T. vaginalis* nuclear protein specifically recognizes wild-type but not mutant M3 sequences. The conservation of the position of M3 relative to the TSS and the effect that mutations within the motif have on transcriptional activity are consistent with this core promoter element being recognized by a *T. vaginalis* transcription factor. As a first step toward identifying such a factor, a *T. vaginalis* nuclear protein was shown to specifically recognize a ³²P-labeled, double-stranded M3-containing probe using EMSAs (Fig. 5A, lane 2). When the wild-type M3 sequence within this probe was mutated at the conserved positions necessary for M3 activity, protein binding was blocked, demonstrating that the protein specifically binds a functional M3 motif (Fig. 5A, lanes 3 to 6). Competition assays likewise showed that three unlabeled probes containing different wild-type M3 motifs are capable of competing for the binding of the protein to the ³²P-labeled wild-type probe (Fig. 5B, lanes 3, 7, 8, and 9). In contrast, unlabeled probes with triple and single point mutations within M3 do not compete (Fig. 5B, lanes 4, 5, and 6). These data demonstrate the presence of a protein(s) in *T. vaginalis* NEs capable of recognizing four consensus M3 motifs.

Isolation of M3BP by DNA affinity chromatography. We have used DNA affinity chromatography to isolate M3BP. Biotinylated wild-type and mutant M3 probes were bound to streptavidin-Sepharose beads, and *T. vaginalis* crude NE was passed over an M3 mutant column to preclear the extract of nonspecific DNA binding proteins. The precleared NE was then applied to either the M3 wild-type or mutant affinity columns. After the columns were extensively washed, proteins were eluted from the columns with increasing concentrations of KCl. The resulting salt fractions were assayed by EMSA for M3 DNA binding activity (Fig. 5D and E). Binding activity was eluted in the 400 mM KCl fraction eluted from the wild-type DNA affinity column (Fig. 5D, lane 8) and was absent from the same fraction eluted from the mutant column (Fig. 5E, lane 8). Both 400 mM KCl fractions were then analyzed by MudPIT mass spectrometry to identify a putative M3BP. Proteins that were represented by at least two peptides were present exclusively in the wild-type sample (data not shown). The most abundant protein identified (TVAG_225940) is a 20.45-kDa protein with a C-terminal Myb-like DNA binding domain comprised of the characteristic R2 and R3 repeats (48, 49).

An alignment of this putative M3BP with the human c-Myb DNA binding domain shows that the 3 critical amino acids of the human c-Myb protein known to make contact with the Myb

expression constructs were transfected into *T. vaginalis*. qRT-PCR was used to measure *CAT* transcript levels. Transcriptional activities are expressed relative to the respective wild-type levels (first sequence in the graphs). *CAT* expression values were normalized to the expression of a neomycin gene, driven by a wild type β -tubulin promoter, on the same expression construct. The M3 and M5 motifs are shown in green and blue, respectively. Mapped TSSs are in boldface, underlined, and denoted by a curved arrow. The frequency of each TSS mapped to a specific nucleotide is indicated next to the arrow. Mutations and insertions introduced into wild-type constructs are shown in red.

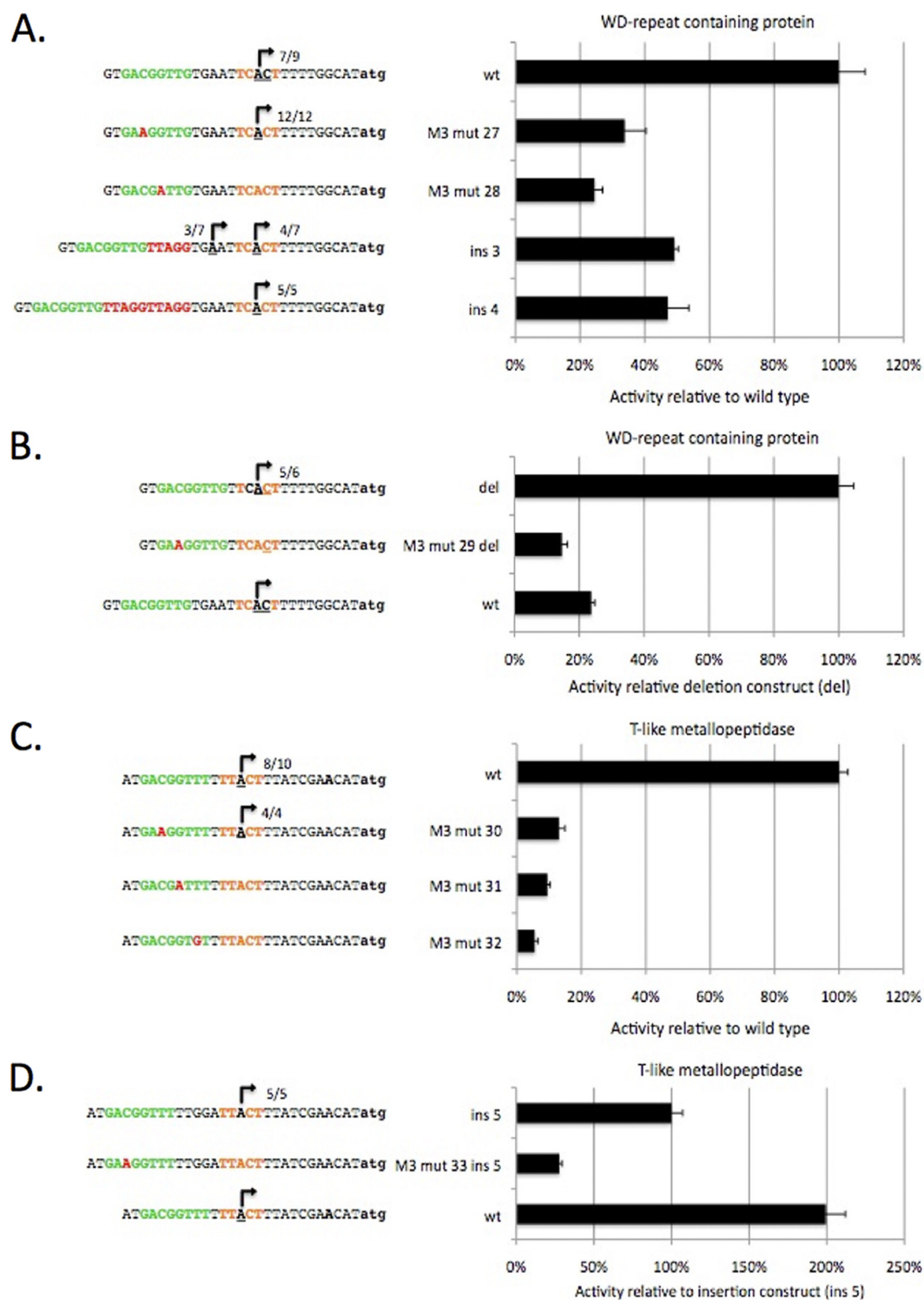


FIG. 4. Transcriptional activity of genes with M3-Inr promoters. The sequence ~500 bp upstream of the WD-repeat-containing protein gene (A and B) and T-like metalloproteinase gene (C and D) were cloned upstream of the *CAT* reporter gene, and the effects of mutations on transcription were assayed as described in the legend to Fig. 3. Motif 3 and Inr elements are shown in green and orange, respectively. Mutations are shown in red, and experimentally determined TSSs are in boldface, underlined, and denoted by curved arrows with the frequency of each TSS indicated.

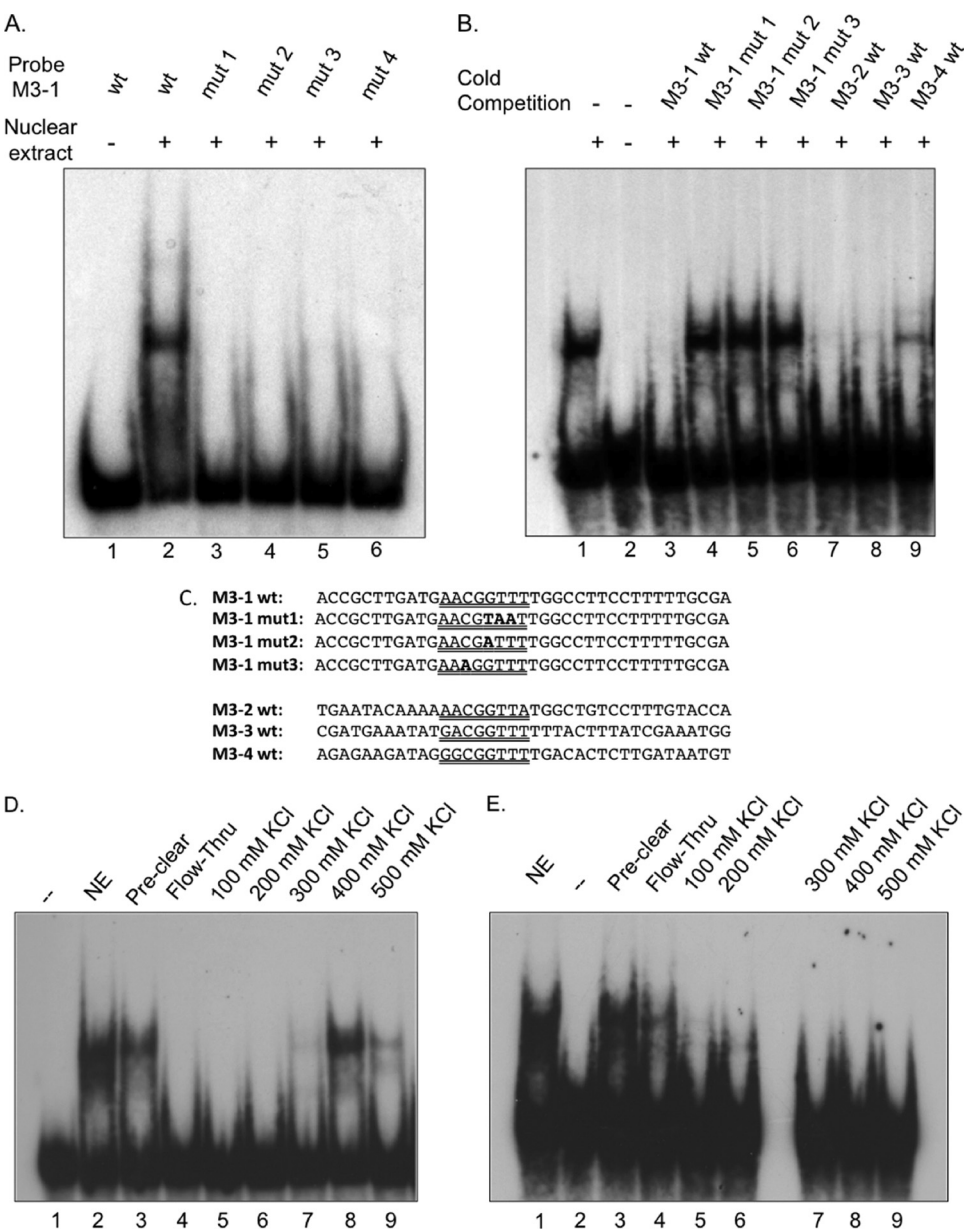


FIG. 5. Affinity isolation of M3-specific DNA binding activity from *T. vaginalis* nuclear extracts. (A) Mobility shifts assays using nuclear extracts and ³²P-labeled M3-1 wild-type and mutant probes shown in panel C; (B) mobility shift assays using a ³²P-labeled M3-1 wild-type (wt) probe and nuclear extracts in the presence of a 100 molar excess of unlabeled probes (listed in panel C); (D) mobility shift assays using ³²P-labeled M3-1 wild-type probe and NE or fractions from wild-type (D) or mutant (E) DNA affinity chromatography.

recognition element (Lys104, Lys158, and Asn159) and the canonical three tryptophans separated by ~19 amino acids in the Myb R2 and R3 repeats are conserved in the putative M3BP (Fig. 6A and C) (48). Except for the conserved C-terminal DNA binding domains, the putative M3BP was found to be a novel protein present only in *T. vaginalis*, as BLASTp analysis of GenBank using the N terminus of the protein reveals no significant similarity with proteins in other organisms. Although there are over 480 Myb-like proteins annotated in the *T. vaginalis* genome (trichodb.org), the three peptides identified by MudPIT mass spectrometry (Fig. 6B) match only the putative M3BP.

Recombinant M3BP specifically recognizes M3 *in vitro* and *in vivo*. To confirm that the putative M3BP has M3-specific DNA binding activity, M3BP was expressed in *Escherichia coli* and the recombinant protein (rM3BP) was purified as a C-terminal histidine-tagged fusion protein. In addition to the 20.45-kDa full-length rM3BP, a less abundant ~17-kDa breakdown product containing the C-terminal His-tagged Myb DNA binding domain was obtained (data not shown). Using this protein preparation, 2 different wild-type M3 probes and 3 different mutant M3 probes were used in EMSA. M3BP was shown to specifically bind the functional and not the nonfunctional M3 motifs (Fig. 7A and B). Two shifted products were observed,

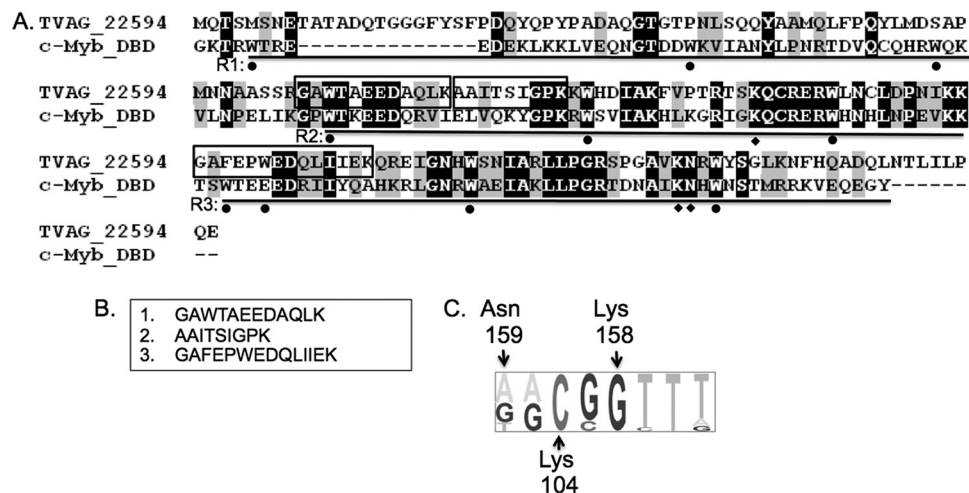


FIG. 6. Sequence analyses of *T. vaginalis* M3BP. (A) Pair-wise alignment of M3BP with human c-Myb (CAF04477) Myb domains. Diamonds, conserved amino acids that make DNA contacts to the MRE; circles, canonical evenly spaced tryptophan residues. The R1, R2, and R3 domains of c-Myb are indicated and underlined. (B) List of the M3BP peptides identified by MudPIT mass spectrometry. (C) The consensus sequence of M3 is shown with arrows pointing to the nucleotides within the metazoan MRE known to contact the Myb DNA binding domain. Contacted amino acid residues (diamonds in panel A) are conserved in M3BP. The residue numbers listed correspond to the human c-Myb sequence.

with the faster-migrating product matching the one observed when NE was used instead of the rM3BP (Fig. 7A and B, lanes 2 and 3). These data indicate that M3BP is cleaved upon purification of NE, consistent with the lack of MudPIT mass spectrometry-identified peptides from the N terminus of the protein (Fig. 6A and B). To further establish the specificity of rM3BP, EMSA experiments were performed using 3 different ³²P-labeled M3 consensus sequence probes and 100 molar excess of either unlabeled wild-type or mutant M3 probes. All tested wild-type M3 probes were shown to compete for the binding of rM3BP to wild-type M3 probes (Fig. 7C, lanes 3, 7, and 11), while no mutant probes were capable of competing (Fig. 7C, lanes 4 to 6, 8 to 10, and 12 to 13). These data definitively establish that the ~20.5-kDa Myb-like DNA binding domain containing M3BP recognizes the M3 core promoter element in *T. vaginalis*.

To directly demonstrate that M3BP binds to M3 *in vivo*, chromatin immunoprecipitation analyses were performed using the IgG fraction purified from polyclonal antisera raised against M3BP. As shown in Fig. 8, the promoter region upstream of the 40S ribosomal protein S15-2 was specifically recovered subsequent to cross-linking of the protein to DNA and precipitation of the DNA with the IgG purified from anti-M3BP antisera (Fig. 8, top panel, anti-M3BP). As a negative control, we tested whether a region of noncoding DNA lacking the M3 motif was present in the anti-M3BP precipitated DNA and found that it was not (Fig. 8, bottom panel, anti-M3BP). We also demonstrated that the promoter region of the 40S ribosomal protein S15-2 was not nonspecifically bound to the columns used to capture the immunoprecipitated DNA by showing that it was not present when antiserum was omitted from the immunoprecipitation (Fig. 8, top panel, no antibody). These data clearly demonstrate that M3BP is capable of binding the M3 motif *in vivo*.

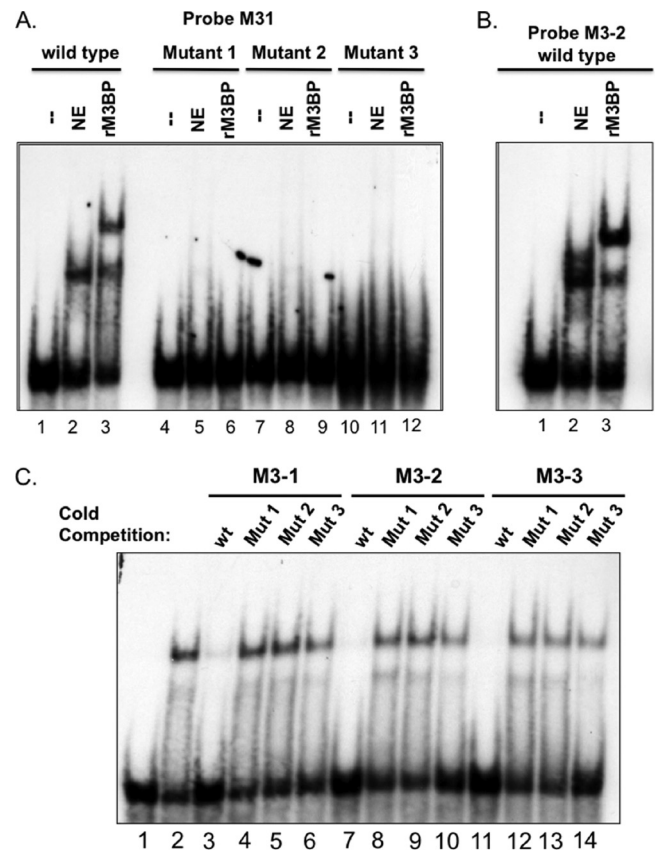


FIG. 7. rM3BP specifically recognizes the motif 3 element *in vitro*. (A) Mobility shift assays using labeled M3-1 wild-type and mutant probes, as indicated, with either 20 µg crude NE or 20 ng of rM3BP; (B) mobility shift assays using labeled M3-2 wild-type probe with either NE or rM3BP; (C) cold competition mobility shift assays using different wild-type labeled M3 probes (M3-1, M3-2, and M3-3), as indicated, and 100× molar excess of unlabeled wild-type (lanes 3, 7, and 11) or mutant M3-1 (lanes 4 to 6, 8 to 11, and 12 to 14) probes. Lane 1, M3-1 probe alone; lane 2, M3-1 probe plus rM3BP without cold probe. The sequences of all probes are shown in Fig. 6C.

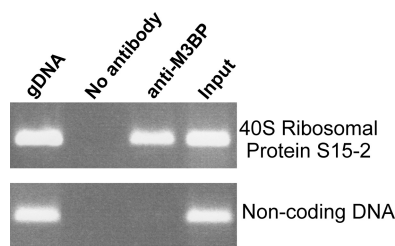


FIG. 8. Endogenous M3BP binding to motif 3 *in vivo*. Chromatin immunoprecipitation analyses were performed using an IgG fraction purified from anti-M3BP antisera to precipitate DNA-protein complexes, followed by PCR using primers upstream and downstream of the M3 motif in the 40S ribosomal protein S15-2 gene (top panel) or primers corresponding to noncoding DNA (bottom panel). The DNA used for PCR is indicated above each lane. gDNA, genomic DNA purified from *T. vaginalis*; no antibody, DNA recovered from protein A columns that do not contain the anti-M3BP IgG; anti-M3BP, DNA recovered from anti-M3BP IgG protein A columns; input, sheared DNA prior to incubation with protein A columns.

DISCUSSION

Little is known about the core promoters that are required to direct transcription of genes in evolutionarily diverse unicellular protists such as *T. vaginalis*. Prior to the work reported here, the only core promoter element known to function in *T. vaginalis* was the Inr element, which is remarkably similar to the metazoan Inr, although initiation of transcription from these related motifs is regulated by unrelated proteins in trichomonads (42, 56) and metazoans (23, 32).

Here, we have identified 2 novel core promoter elements by searching a nonredundant data set of 5' UTRs composed of expressed *T. vaginalis* genes for overrepresented DNA motifs. In addition to detecting the previously described Inr (41), M3 and M5 were identified. Although the parameters of our original search did not detect a third promoter motif, we did identify an additional motif during our analyses of the TSSs from target promoters. This third motif, GG(T/C)(T/C), is preferentially located upstream of the M5 motif. Mutational and functional analyses of these elements demonstrate that they have the properties of core promoter elements: each is present in multiple genes located a fixed distance relative to the TSS, and mutation of the elements reduces transcriptional activity and/or alters TSS selection.

Approximately 92% of the 5' UTRs examined in this study contain an M3 [(A/G/T)(A/G)C(G/C)G(T/C)T(T/A/G)], M5 (CCTTT), or Inr [(T/C/A)(C/T)A₍₊₁₎(T/C/A)(T/A)] motif, either singularly or in combination. The majority of these (82%) contained only an Inr, consistent with previous predictions (7) that most *T. vaginalis* protein-coding genes utilize the Inr to direct transcription initiation. Interestingly, the methods applied in this study indicated that only 4% of 5' UTRs containing an Inr motif also contained an M3 element, and no additional potential core promoter elements were identified in these UTRs. Similarly, the M5 motif also appears to be capable of functioning independently of M3 or any other detectable core promoter, as ~32% of the 5' UTRs examined that contain an M5 motif appear to have no other conserved promoter motif. In contrast, roughly 90% of genes that contain the M3 motif also have either an M5 or Inr motif upstream of the

translation start site, indicating that M3 functions almost exclusively in synergy with one of these 2 initiator elements.

M3 and M5 do not fit the consensus sequence of any known eukaryotic core promoter element of protein-coding genes. The M5 motif (CCTTT) was shown to be an alternative initiator element, as it surrounds and contains the TSS. It is located almost exclusively 5 to 15 bp upstream of the translation start codon ATG, the predominant position of the Inr (41, 54). M5 and Inr appear to be mutually exclusive; protein-coding genes use either one or the other of these elements to mark the TSS. Interestingly, the M5 consensus sequence, CCTTT, mostly precedes ribosomal protein genes. It is unclear why an alternative initiator element such as M5 would have evolved as a core promoter specific to a single family of genes. An expanded family of 422 genes encodes ribosomal proteins in *T. vaginalis*, making it possible that the large number of genes together with the vital importance of translation led to the selection of an exclusive core promoter for this class of genes.

The M3 motif, which primarily resides at a fixed distance upstream of either an M5 or an Inr motif, plays a critical role in directing the TSS to the second C of the M5 motif. Mutation of the M3 element by altering either its sequence composition or its distance relative to M5 all but abolishes transcription in promoters that contain both an M3 and an M5. M3 was also shown to work synergistically to direct transcription initiating in the Inr; however, mutations within M3 do not abolish Inr-driven transcription, as the Inr is capable of directing transcription independent of M3, albeit at lower levels.

It is notable that Cong et al. have identified by bioinformatic analyses the presence of a conserved motif upstream of *T. vaginalis* histone genes which is the inverse of the M3 motif (11). This suggests that M3 (and their motif, which was referred to as motif 2) may function in either orientation. Most genes appear to use the M3-M5 and M3-Inr orientation described here, while the histone gene family (72 genes) may utilize M3 in a reverse orientation.

The TSS of the M5 motif is also unusual, as transcription initiates at a cytosine [CC₍₊₁₎TTT] instead of a canonical purine (Pu). Eukaryotic TSSs typically contain a dinucleotide Py-Pu₍₊₁₎, as found in the *T. vaginalis* Inr element, (T/C/A)C A₍₊₁₎N(T/A) (8, 32, 41). This Py-Pu dinucleotide has been shown to act as a stronger promoter and to be preferred over a Py-Py start site in a classic Inr (8, 41). The mechanism by which a C is selected as the TSS in M5 is yet to be determined.

With the exception of MTE and XCPE2, all known metazoan core promoter elements could be found in the 60 bp immediately upstream of *T. vaginalis* protein-coding genes when a direct search for these elements was conducted. However, the detected motifs are present at a frequency greater than that predicted by random chance, and none of them, besides the Inr, had strong positional conservation within the *T. vaginalis* 5' UTRs. These data indicate that, with the exception of the Inr, metazoan core promoter elements are not used by *T. vaginalis*. This is somewhat unexpected, as most of the core transcription factors of the metazoan PIC, such as TFIIB, the majority of TAFs that comprise the TFIID complex, TFIIE, and all but one subunit of RNA polymerase II, are encoded in the *T. vaginalis* genome (7). It should be noted, however, that our data do not preclude the possibility that a subset of promoters may use one or more of the previously

described metazoan core promoter elements to direct transcription initiation. Alternatively, the conserved PIC proteins may function to drive transcription from either the *T. vaginalis* Inr or M5 promoter, interacting either directly or indirectly with IBP39 (36, 42, 56) and/or M3BP, discussed below.

The M3 consensus sequence, (A/G/T)(A/G)C(G/C)G(T/C)T(T/A/G), strongly resembles the classic eukaryotic MRE, Py-AAC(T/G)G, which is found in distal promoter regions of genes in a broad range of eukaryotes and acts to regulate an array of developmental processes (19, 30, 49). For example, in humans, the MRE is recognized by a family of Myb proteins (c-Myb, A-Myb, and B-Myb) that regulate processes such as hematopoiesis and spermatogenesis (38, 49). MRE sequences in plants are involved in regulating multiple processes, including embryogenesis, flower morphology, light responsiveness, and response to mechanical damage (17, 44, 45). MREs have also been described in parasitic protists to function as distal promoter elements. The transcription of several encystation-specific genes of *Giardia lamblia* is regulated by a MRE, located 18 to 40 bp upstream of the TSSs (62). The *Entamoeba histolytica* Myb-dr (EhMyb-dr) protein, which recognizes a C-rich upstream motif, also regulates genes preferentially expressed during *E. histolytica* encystation (13). Within the *T. vaginalis* genome, the distal promoter of the malic enzyme gene contains two divergent MRE sequences, MRE-1/MRE-2r (TAACGATA) and MRE-2f (TATCGT), recognized by three Myb proteins, Myb1, Myb2, and Myb3 (28, 51). Similar to MRE motifs in other eukaryotes, *T. vaginalis* MRE-1 (TvMRE-1)/MRE-2r are involved in regulating transcription in response to environmental conditions (specifically, iron concentration). However, TvMRE-1/MRE-2r do not contain the classic eukaryotic MRE core motif, CNGTT.

Demonstrating that an MRE-like element can function as a core promoter element preceding protein-coding genes, as shown here for M3, shows that it represents a novel function for the MRE. It is notable that the location of the M3 element relative to the TSS and its apparent role in selecting the TSS make it analogous to the TATA box (8, 53), a common eukaryotic core promoter element that *T. vaginalis* does not appear to use. These observations support the proposal that the M3 motif has replaced the TATA box in this lineage of divergent eukaryotes.

The *T. vaginalis* transcription factor that specifically recognizes the M3 promoter was identified using DNA affinity chromatography. This protein, called M3BP, was shown to directly interact with the M3 motif both *in vitro* and *in vivo*. M3BP has a novel N terminus that bears no resemblance to any known protein; however, it contains an R2-R3 Myb DNA binding domain at its C terminus. The ~20.5-kDa M3BP is ~23 to 30% identical and 35 to 43% similar to *T. vaginalis* Myb1, Myb2, and Myb3 (28, 51), with the only significant similarity being found in the R2-R3 Myb domains. Although there are ~480 Myb-like proteins encoded in the *T. vaginalis* genome, M3BP was the only Myb-like protein identified using sequence-specific DNA affinity chromatography. Perhaps the unique N terminus of M3BP distinguishes it from other Myb-like proteins, thus allowing it to specifically recognize M3. It is also possible that other *T. vaginalis* Myb-like proteins recognize M3 with various affinities, which did not lead to their detection, as multiple

proteins are thought to interact with core promoter elements, such as the metazoan Inr element and TATA box (4, 6, 26).

M3BP is the first Myb-like protein implicated in directing basal transcription of protein-coding genes. Another Myb-like protein, SNAP190, has been shown to interact with the RNA polymerase II and RNA polymerase III transcription machinery that transcribes the human U1-U5 and U6 small nuclear RNA (snRNA) genes, respectively. These genes contain a proximal sequence element (PSE) that is sufficient to direct basal transcription (59) and that SNAP190 directly contacts (27, 29, 71). M3BP and SNAP190 have no significant sequence similarities other than at the Myb domains, and BLAST analyses do not indicate these proteins to be homologous. Further studies will be required to determine whether M3BP interacts directly with components of the RNA polymerase II preinitiation complex. It appears that overexpression of M3BP in *T. vaginalis* is lethal, as numerous attempts to obtain transfectants expressing this protein, using vectors and protocols that allow the overexpression of many other *T. vaginalis* genes, failed.

The identification of novel core promoter elements in this divergent eukaryote implies the presence of potentially novel mechanisms underlying gene expression. Characterization of additional *T. vaginalis* proteins that may direct basal-level transcription through interactions with these elements should broaden our understanding of the varied mechanisms that control gene expression in eukaryotes, as instructed through the composition and architecture of the core promoter.

ACKNOWLEDGMENTS

We thank Jennifer Gordon for critical comments on the manuscript and our colleagues in the lab for helpful discussions.

This work was funded by National Institutes of Health R01 grant AI30537 (to P.J.J.) and the Jonsson Cancer Center at UCLA (J.A.W.). A.J.S. was supported by a National Institutes of Health training grant (T32-AI-007323) and a National Research Service Award (F31AI68621).

REFERENCES

1. Anish, R., M. B. Hossain, R. H. Jacobson, and S. Takada. 2009. Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCPE2 and analysis of factor requirements. *PLoS One* 4:e5103.
2. Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2:28–36.
3. Baldauf, S. L. 2003. The deep roots of eukaryotes. *Science* 300:1703–1706.
4. Bártfai, R. R., et al. 2004. TBP2, a vertebrate-specific member of the TBP family, is required in embryonic development of zebrafish. *Curr. Biol.* 14: 593–598.
5. Burke, T. W., P. J. Willy, A. K. Kutach, J. E. Butler, and J. T. Kadonaga. 1998. The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box. *Cold Spring Harbor Symp. Quant. Biol.* 63:75–82.
6. Carcamo, J., L. Buckbinder, and D. Reinberg. 1991. The initiator directs the assembly of a transcription factor IID-dependent transcription complex. *Proc. Natl. Acad. Sci. U. S. A.* 88:8052–8056.
7. Carlton, J. M., et al. 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315:207–212.
8. Carninci, P., et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38:626–635.
9. Cavalier-Smith, T. 1993. Kingdom protozoa and its 18 phyla. *Microbiol. Mol. Biol. Rev.* 57:953–994.
10. Chalkley, G. E., and C. P. Verrijzer. 1999. DNA binding site selection by RNA polymerase II TAFs: a TAFII250-TAFII150 complex. *EMBO J.* 18: 4835–4845.
11. Cong, P., Y. Luo, W. Bao, and S. Hu. 2010. Genomic organization and promoter analysis of the *Trichomonas vaginalis* core histone gene families. *Parasitol. Int.* 59:29–34.
12. Delgadillo, M. G., D. R. Liston, K. Niazi, and P. J. Johnson. 1997. Transient

- and selectable transformation of the parasitic protist *Trichomonas vaginalis*. *Proc. Natl. Acad. Sci. U. S. A.* **94**:4716–4720.
13. Ehrenkaufer, G. M., J. A. Hackney, and U. Singh. 2009. A developmentally regulated Myb domain protein regulates expression of a subset of stage-specific genes in *Entamoeba histolytica*. *Cell. Microbiol.* **11**:898–910.
 14. Elias, J. E., and S. P. Gygi. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**:207–214.
 15. Eng, J. K., A. L. McCormack, and J. R. Yates III. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**:976–989.
 16. Evans, R., J. A. Fairley, and S. G. Roberts. 2001. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes Dev.* **15**:2945–2949.
 17. Feldbrugge, M., M. Sprenger, K. Hahlbrock, and B. Weisshaar. 1997. PcMYB1, a novel plant protein containing a DNA-binding domain with one MYB repeat, interacts in vivo with a light-regulatory promoter unit. *Plant J.* **11**:1079–1093.
 18. Florens, L., et al. 2006. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**:303–311.
 19. Ganter, B., S. T. Chao, and J. S. Lipsick. 1999. Transcriptional activation by the Myb proteins requires a specific local promoter structure. *FEBS Lett.* **460**:401–410.
 20. Gardner, W. A., Jr., D. E. Culbertson, and B. D. Bennett. 1986. *Trichomonas vaginalis* in the prostate gland. *Arch. Pathol. Lab Med.* **110**:430–432.
 21. Goldberg, M. 1979. Ph.D. dissertation. Stanford University, Stanford, CA.
 22. Grillo, G., M. Attimonelli, S. Liuni, and G. Pesole. 1996. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. *Comput. Appl. Biosci.* **12**:1–8.
 23. Gross, P., and T. Oelgeschlager. 2006. Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.* **73**:225–236.
 24. Gunderson, J., et al. 1995. Phylogeny of trichomonads inferred from small-subunit rRNA sequences. *J. Eukaryot. Microbiol.* **42**:411–415.
 25. Hampsey, M. 1998. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.* **62**:465–503.
 26. Hansen, S. K., S. Takada, R. H. Jacobson, J. T. Lis, and R. Tjian. 1997. Transcription properties of a cell type-specific TATA-binding protein, TRF. *Cell* **91**:71–83.
 27. Hinkley, C. S., H. A. Hirsch, L. Gu, B. LaMere, and R. W. Henry. 2003. The small nuclear RNA-activating protein 190 Myb DNA binding domain stimulates TATA box-binding protein-TATA box recognition. *J. Biol. Chem.* **278**:18649–18657.
 28. Hsu, H.-M., S.-J. Ong, M.-C. Lee, and J.-H. Tai. 2009. Transcriptional regulation of an iron-inducible gene by differential and alternate promoter entries of multiple Myb proteins in the protozoan parasite *Trichomonas vaginalis*. *Eukaryot. Cell* **8**:362–372.
 29. Hung, K. H., M. Titus, S. C. Chiang, and W. E. Stumph. 2009. A map of *Drosophila melanogaster* small nuclear RNA-activating protein complex (DmSNAPc) domains involved in subunit assembly and DNA binding. *J. Biol. Chem.* **284**:22568–22579.
 30. Jin, H., and C. Martin. 1999. Multifunctionality and diversity within the plant MYB-gene family. *Plant Mol. Biol.* **41**:577–585.
 31. Juven-Gershon, T., J. Y. Hsu, and J. T. Kadonaga. 2006. Perspectives on the RNA polymerase II core promoter. *Biochem. Soc. Trans.* **34**:1047–1050.
 32. Juven-Gershon, T., J. Y. Hsu, J. W. Theisen, and J. T. Kadonaga. 2008. The RNA polymerase II core promoter—the gateway to transcription. *Curr. Opin. Cell Biol.* **20**:253–259.
 33. Kadonaga, J. T. 2002. The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.* **34**:259–264.
 34. Keeling, P. J., et al. 2005. The tree of eukaryotes. *Trends Ecol. Evol.* **20**:670–676.
 35. Keeling, P. J., and J. D. Palmer. 2000. Phylogeny: Parabasal flagellates are ancient eukaryotes. *Nature* **405**:635–637.
 36. Lau, A. O. T., A. J. Smith, M. T. Brown, and P. J. Johnson. 2006. *Trichomonas vaginalis* initiator binding protein (IBP39) and RNA polymerase II large subunit carboxy terminal domain interaction. *Mol. Biochem. Parasitol.* **150**:56–62.
 37. Lee, D. H., et al. 2005. Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol. Cell. Biol.* **25**:9674–9686.
 38. Lei, W., J. J. Rushton, L. M. Davis, F. Liu, and S. A. Ness. 2004. Positive and negative determinants of target gene specificity in Myb transcription factors. *J. Biol. Chem.* **279**:29519–29527.
 39. Lemon, B., and R. Tjian. 2000. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* **14**:2551–2569.
 40. Liston, D. R., J.-C. Carrero, and P. J. Johnson. 1999. Upstream regulatory sequences required for expression of the *Trichomonas vaginalis* [alpha]-succinyl CoA synthetase gene. *Mol. Biochem. Parasitol.* **104**:323–329.
 41. Liston, D. R., and P. J. Johnson. 1999. Analysis of a ubiquitous promoter element in a primitive eukaryote: early evolution of the initiator element. *Mol. Cell. Biol.* **19**:2380–2388.
 42. Liston, D. R., A. O. T. Lau, D. Ortiz, S. T. Smale, and P. J. Johnson. 2001. Initiator recognition in a primitive eukaryote: IBP39, an initiator-binding protein from *Trichomonas vaginalis*. *Mol. Cell. Biol.* **21**:7872–7882.
 43. Magnus, M., R. Clark, L. Myers, T. Farley, and P. J. Kissinger. 2003. *Trichomonas vaginalis* among HIV-infected women: are immune status or protease inhibitor use associated with subsequent *T. vaginalis* positivity? *Sex. Transm. Dis.* **30**:839–843.
 44. Mahjoub, A., et al. 2009. Overexpression of a grapevine R2R3-MYB factor in tomato affects vegetative development, flower morphology and flavonoid and terpenoid metabolism. *Plant Physiol. Biochem.* **47**:551–561.
 45. Mellway, R. D., L. T. Tran, M. B. Prouse, M. M. Campbell, and C. P. Constabel. 2009. The wound-, pathogen-, and UV-B-responsive MYB134 gene encodes an R2R3 MYB transcription factor that regulates proanthocyanidin synthesis in poplar. *Plant Physiol.* **150**:924–941.
 46. Meng, Z., et al. 2006. Probing early growth response 1 interacting proteins at the active promoter in osteoblast cells using oligoprecipitation and mass spectrometry. *J. Proteome Res.* **5**:1931–1939.
 47. Miller, M., Y. Liao, A. M. Gomez, C. A. Gaydos, and D. D'Mellow. 2008. Factors associated with the prevalence and incidence of *Trichomonas vaginalis* infection among African American women in New York City who use drugs. *J. Infect. Dis.* **197**:503–509.
 48. Ogata, K., et al. 1994. Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79**:639–648.
 49. Oh, I. H., and E. P. Reddy. 1999. The myb gene family in cell growth, differentiation and apoptosis. *Oncogene* **18**:3017–3033.
 50. Ohler, U., G. C. Liao, H. Niemann, and G. M. Rubin. 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**:RESEARCH0087.
 51. Ong, S.-J., H.-M. Hsu, H.-W. Liu, C.-H. Chu, and J.-H. Tai. 2006. Multifarious transcriptional regulation of adhesion protein gene ap65-1 by a novel Myb1 protein in the protozoan parasite *Trichomonas vaginalis*. *Eukaryot. Cell* **5**:391–399.
 52. Petrin, D., K. Delgaty, R. Bhatt, and G. Garber. 1998. Clinical and microbiological aspects of *Trichomonas vaginalis*. *Clin. Microbiol. Rev.* **11**:300–317.
 53. Ponjavic, J., et al. 2006. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.* **7**:R78.
 54. Quon, D. V., M. G. Delgadillo, A. Khachi, S. T. Smale, and P. J. Johnson. 1994. Similarity between a ubiquitous promoter element in an ancient eukaryote and mammalian initiator elements. *Proc. Natl. Acad. Sci. U. S. A.* **91**:4579–4583.
 55. Roeder, R. G. 2005. Transcriptional regulation and the role of diverse co-activators in animal cells. *FEBS Lett.* **579**:909–915.
 56. Schumacher, M. A., A. O. T. Lau, and P. J. Johnson. 2003. Structural basis of core promoter recognition in a primitive eukaryote. *Cell* **115**:413–424.
 57. Schwelke, J. R., and D. Burgess. 2004. Trichomoniasis. *Clin. Microbiol. Rev.* **17**:794–803.
 58. Smale, S. T., and D. Baltimore. 1989. The “initiator” as a transcription control element. *Cell* **57**:103–113.
 59. Smale, S. T., and J. T. Kadonaga. 2003. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**:449.
 60. Sorvillo, F., L. Smith, P. Kerndt, and L. Ash. 2001. *Trichomonas vaginalis*, HIV, and African-Americans. *Emerg. Infect. Dis.* **7**:927–932.
 61. Stark, J. R., et al. 2009. Prospective study of *Trichomonas vaginalis* infection and prostate cancer incidence and mortality: Physicians' Health Study. *J. Natl. Cancer Inst.* **101**:1406–1411.
 62. Sun, C. H., D. Palm, A. G. McArthur, S. G. Svard, and F. D. Gillin. 2002. A novel Myb-related protein involved in transcriptional activation of encystation genes in *Giardia lamblia*. *Mol. Microbiol.* **46**:971–984.
 63. Sutcliffe, S., et al. 2006. Plasma antibodies against *Trichomonas vaginalis* and subsequent risk of prostate cancer. *Cancer Epidemiol. Biomarkers Prev.* **15**:939–945.
 64. Tabb, D. L., W. H. McDonald, and J. R. Yates III. 2002. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1**:21–26.
 65. Thomas, M. C., and C. M. Chiang. 2006. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**:105–178.
 66. Tokusumi, Y., Y. Ma, X. Song, R. H. Jacobson, and S. Takada. 2007. The new core promoter element XCEP1 (X core promoter element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Mol. Cell. Biol.* **27**:1844–1858.
 67. Van Der Pol, B., et al. 2008. *Trichomonas vaginalis* infection and human immunodeficiency virus acquisition in African women. *J. Infect. Dis.* **197**:548–554.
 68. Washburn, M. P., D. Wolters, and J. R. Yates III. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**:242–247.
 69. Weinstock, H., S. Berman, and W. Cates, Jr. 2004. Sexually transmitted diseases among American youth: incidence and prevalence estimates, 2000. *Perspect. Sex. Reprod. Health* **36**:6–10.

70. **Wohlschlegel, J. A.** 2009. Identification of SUMO-conjugated proteins and their SUMO attachment sites using proteomic mass spectrometry. *Methods Mol. Biol.* **497**:33–49.
71. **Wong, M. W., et al.** 1998. The large subunit of basal transcription factor SNAPc is a Myb domain protein that interacts with Oct-1. *Mol. Cell. Biol.* **18**:368–377.
72. **World Health Organization.** 1996. Sexually transmitted diseases fact sheet no. 110. World Health Organization, Geneva, Switzerland.
73. **Yakovchuk, P., E. Protozanova, and M. D. Frank-Kamenetskii.** 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.* **34**:564–574.
74. **Yap, E. H., et al.** 1995. Serum antibodies to *Trichomonas vaginalis* in invasive cervical cancer patients. *Genitourin. Med.* **71**:402–404.
75. **Zhang, Z. F., and C. B. Begg.** 1994. Is *Trichomonas vaginalis* a cause of cervical neoplasia? Results from a combined analysis of 24 studies. *Int. J. Epidemiol.* **23**:682–690.